Measuring Mortality Information in Clinical Data Warehouses

Barrett Jones[1], David K. Vawdrey, PhD[2]

[1]Department of Statistics, Brigham Young University, Provo, UT
[2] Department of Biomedical Informatics, Columbia University, New York, NY

Abstract

The ability to track and report long-term outcomes, especially mortality, is essential for advancing clinical research. The purpose of this study was to present a framework for assessing the quality of mortality information in clinical research databases. Using the clinical data warehouse (CDW) at Columbia University Medical Center as a case study, we measured: 1) agreement in vital status between our institution's patient registration system and the U.S. Social Security Administration's Death Master File (DMF), 2) the proportion of patients marked as deceased according to the DMF records who had subsequent visits to our institution, and 3) the proportion of patients still living according to Columbia's CDW who were over 100 and 120 years of age. Of 33,295 deaths recorded in our institution's patient registration system, 13,167 (39.5%) did not exist in the DMF. Of 315,037 patients in our CDW who marked as deceased according to the DMF, 2.1% had a subsequent clinical encounter at our institution. The proportion of patients still living according to Columbia's CDW who were over 100 and 120 years of age was 43.6% and 43.1%, respectively. These measures may be useful to other clinical research investigators seeking to assess the quality of mortality data (1-4).

Introduction

The ability to track and report long-term outcomes is essential for advancing clinical research. Whether a subject is living or deceased is perhaps the most obvious and seemingly easy-to-measure outcome; however, many clinical research databases either lack patient mortality information or require extensive effort to collect this information and keep it up-to-date.

In the United States, there are two national resources that provide mortality information: the Social Security Administration's Death Master File (DMF), and the Centers for Disease Control's National Death Index (NDI). The DMF is updated weekly and is available as a bulk file that can be purchased at minimal cost (approximately $1,825, plus a fee for ongoing updates). The NDI is updated yearly and must be queried one name at a time, at a cost of $0.21 per name. Because of its lower bulk cost and timeliness of information, the DMF is more suitable for integration into institutional clinical data warehouses. However, unlike the NDI, the DMF does not contain cause-of-death information. Moreover, the completeness of data in the DMF is decreasing: in 2011, because of concerns about state vs. federal ownership of certain records in the DMF amidst public concerns over identity theft, the Social Security Administration expunged approximately 5% of the DMF's 89 million records and announced that about one million fewer records would be added each year

Our institution, Columbia University Medical Center (CUMC), has a clinical data warehouse (CDW) containing information for over 4.5 million patients dating to 1988. The CDW contains death information from both the hospital's patient registration system as well as from the DMF. The quality of death information in institutional clinical research databases—and particularly the impact of the 2011 modifications to the DMF—has not been carefully examined (1,2). The purpose of this study was to present a framework for assessing the quality of mortality information in clinical research databases, using the CDW at CUMC as a case study.

Background

Collection of mortality data in the United States is a complex process (Figure 1). The U.S. Social Security Administration (SSA) began collecting data on deaths in 1962. SSA receives death data primarily from state reporting, families of the deceased, funeral homes and financial institutions.

As a result of a 1978 court case under the Freedom of Information Act, SSA is required to release its death information to the public. Known as the Death Master File (DMF), the publicly available information concerning the deceased includes: first and last name, date of birth, month and year of death, social security number, and whether the death has been verified or a death certificate has been observed. Prior to 2011, when the SSA removed state records from the DMF, the file also contained the last ZIP code of the person while living, and the ZIP code to which any lump sum death benefit was sent (5). The SSA delivers the DMF for public distribution through the National Technical Information Service within the Department of Commerce.

The DMF information has been useful to researchers in many fields, including healthcare delivery, clinical research, and financial services, where it is used to prevent identity theft and tax fraud. Not surprisingly, because the DMF contains a vast number of social security numbers and is publicly available, the resource itself may be considered a risk in terms of facilitating identity theft and tax fraud. Due to these risks, the SSA determined in 2011 that records received from the states could not be redistributed. Approximately 5% of the DMF's 89 million records were removed, and since 2011, about 1 million fewer records have been added each year. Legislation was passed in 2013 to further restrict access to the DMF, requiring individuals and institutions to provide a legitimate business purpose or fraud prevention interest to receive DMF data (6).

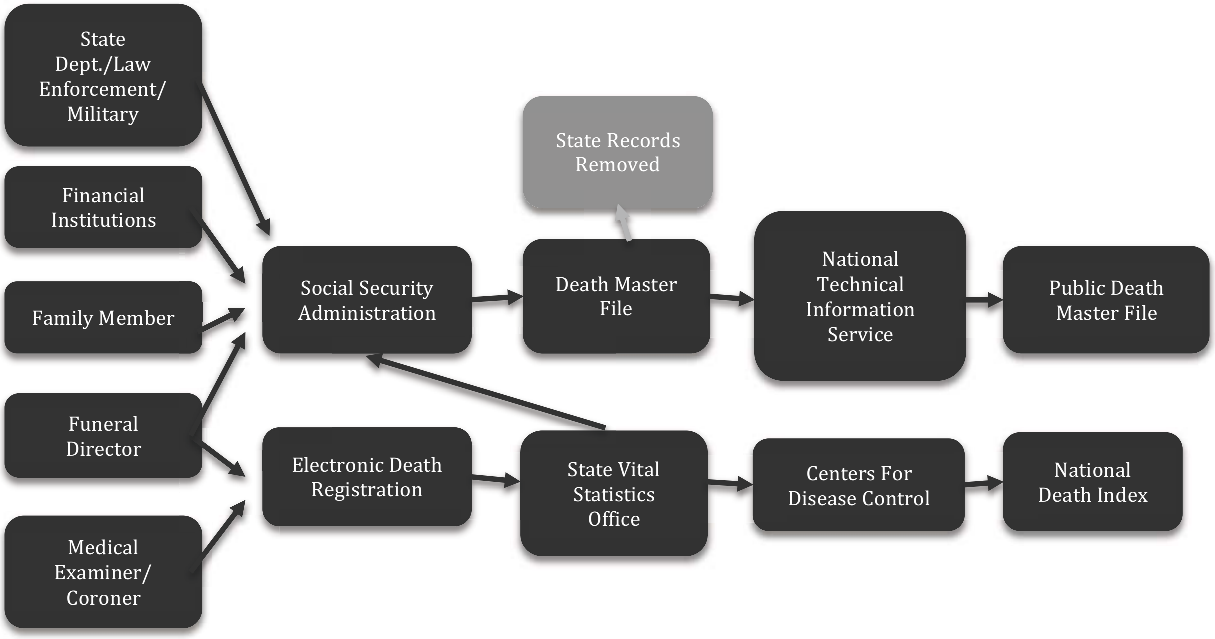
State Dept./Law Enforcement/ Military
Financial Institutions
Family Member
Funeral Director
Medical Examiner/ Coroner
Social Security Administration
Electronic Death Registration
State Records Removed
Death Master File
State Vital Statistics Office
National Technical Information Service
Centers For Disease Control
Public Death Master File
National Death Index

Figure 1. Data collection process for death information.

The Center for Disease Control has access to all state death information, with which they create the National Death Index. The National Death Index is the most accurate source for death information. Currently, it contains death information from the years 1979-2011. The NDI can be queried at a cost of $0.21 per name ($0.15 if cause of death information is excluded). This information is only available to medical researchers who complete the National Center for Health Statistics certification program. This program takes 2-3 months to complete and requires the recipient to have an Institutional Review Board-approved study. In reaction to the restrictions to the DMF, the NDI has attempted to improve the timeliness of data availability, making data available six months after the end of the year without cause of death information. For example, death data for 2013 were made available in June 2014, and the cause of death information should be available near the end of 2014. The dataset from 2012 had some issues and is not currently available, but should be soon in its entirety (7).

Methods

The CDW was queried to find the following: the number of deaths per year 1870-2013, the number of births per year from 1870-2013, the number of patients dead that were born each year from 1870-2013, deaths according to the hospitals registration system, deaths according to the DMF, and the number of patients who had a visit at least one month after their date of death. From these queries we measured: 1) agreement in vital status between our institution's patient registration system and the DMF. Agreement was calculated by summing deaths from the DMF and deaths from the hospital registration system then subtracting the total number of deaths to find the intersection. This number was divided by deaths from the hospital registration system. Agreement was also calculated for each year 1988-2013. 2) The proportion of patients marked as deceased according to the DMF records who had subsequent visits to our institution was calculated for patients with visits one day after date of death, and at least 30 days after death. 3) The proportion of patients still living according to the CDW records who were over 100 and 120 years of age. To further understand the data we also queried the total number of patients in the CDW through December 31, 2013, the number of deceased patients, deceased from the DMF, and deceased from the institution's registration system.

Results

We found an agreement of 60.5% between our institution's registration system and the DMF, this was broken down by year, and it was found that there was significantly less agreement after 2011. There were 6,726 patients marked as deceased that had a subsequent visit to our institution, making up 2.1% of deaths according to the DMF. The SSA often is not supplied with the exact day of death, so to adjust for this the CDW was also queried to find the number of patients who had a visit at least 30 days after their date of death, the result was 6,290 (1.9%). 43.6% of patients 100 years or older were still living according to the CDW, and that number drops to 43.1% for patients over 120.

Table 1. Death information recorded in Columbia's clinical data warehouse

Total patients	4,449,684
Deceased	328,204 (7.4%)
Death Master File (DMF)	315,037
Registration system	33,295
Deaths in registration system not found in DMF*	13,167 (39.5%)
Patients > 100 years	126,601
Living	55,198 (43.6%)
Patients > 120 years	8,644
Living	3,725 (43.1%)
Patients with clinical visit subsequent to date of death in DMF	6,726 (2.1% of 315,037 DMF deaths)

*As of March 2014 DMF update

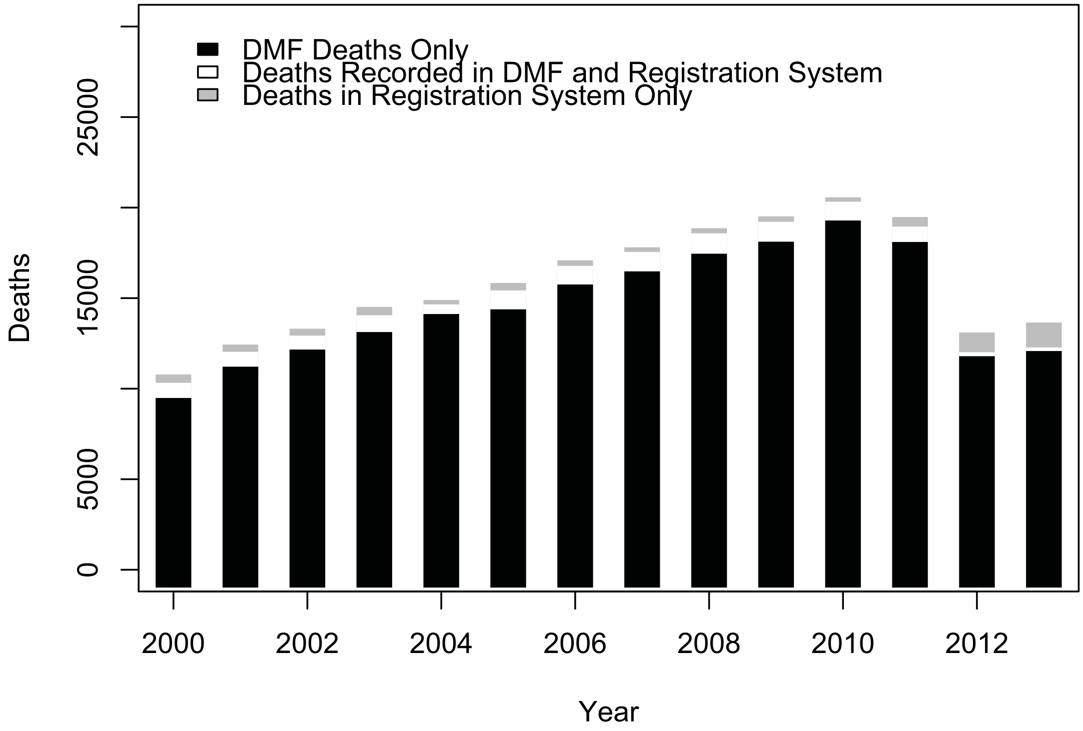

Figure 2. Deaths recorded in Columbia's clinical data warehouse by year (2000-2013).

Discussion

Without integrating data from outside sources, the mortality information in a healthcare delivery organization's clinical data warehouse may be extremely incomplete. We found that agreement and accuracy of the data were low, suggesting a lack of quality in the DMF data for the population our institution serves. If the DMF had perfect agreement we would expect 100% overlap between the deaths from the DMF and deaths recorded in the hospital. We found only 60.5% agreement, meaning only 60.5% of the deaths in the hospital were recorded in the DMF. This finding may be partially explained by the hospital receiving significantly fewer records after 2011, as can be seen in the year-by-year breakdown of deaths in Figure 2. In 2012 and 2013, our data warehouse contained significantly fewer total deaths, and agreement plummeted to 17% and 13% agreement respectively. However, even if 2012 and 2013 data are excluded, we observed only 62% agreement overall. This result is lower than what has been reported in other studies. Hauser et al. found 82% agreement between the DMF and a cohort of known decedents (8). Huntington et al. found 94.7% agreement between the DMF and a random sample of Ohio deaths (9). Newman et al. found 96.5% agreement between the DMF supplemented with state death records and in hospital deaths at a California hospital. Other studies have found similar results (10–14).

Why are the results lower at our institution than others? It must be taken into consideration the size of our CDW, and that the time period covered is greater than these studies. CUMC also serves a predominantly minority population with a large amount of immigrants; these populations are known to have lower representation in the DMF (9,12).

What can be done to improve mortality data quality? Unfortunately there is no simple solution. We would recommend institutions follow the framework laid out in this paper to assess their data quality. Institutions that would like to obtain the highest data quality should use the National Death Index supplemented with the DMF until the full NDI becomes available each year. This is an expensive option. There were 4,121,480 patients currently living according to the CDW. A search of the NDI would cost $865,510.80 for just one year searched with cause of death, and would still cost $618,222.00 without the cause of death included. Access to the SSDI will add at the minimum $1,825 depending on which service is chosen. The other option would be to continue using the DMF despite its shortcomings. An algorithm could potentially be developed to estimate the true number deceased patients in the CDW (15). Neither of these options serve medical researchers well, so it is vital that all who are in this field understand that the only real solution is new legislation (2) to either make the NDI more affordable and available in a timely manner by providing government funding, or to ease restrictions on the DMF and make all records, regardless of where they were obtained, accessible to medical researchers.

References

1. 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 [Internet]. 2013 Jan 1 [cited 2014 Jun 30];6(1):124–8. Available from: http://www.ncbi.nlm.nih.gov/pubmed/23322808

2. Blackstone EH. Demise of a vital resource. J Thorac Cardiovasc Surg [Internet]. 2012 Jan [cited 2014 Jun 30];143(1):37–8. Available from: http://www.sciencedirect.com/science/article/pii/S002252231101302X

3. SSA DMF [Internet]. [cited 2014 Jul 2]. Available from: http://www.ntis.gov/products/ssa-dmf.aspx

4. Social Security Death Record Limits Hamper Researchers - NYTimes.com [Internet]. [cited 2014 Jul 28]. Available from: http://www.nytimes.com/2012/10/09/us/social-security-death-record-limits-hinder-researchers.html?_r=1&

5. Cambridge Journals Online - Cardiology in the Young - Fulltext - Empowering a database with national long-term data about mortality: the use of national death registries . [cited 2014 Jul 1]; Available from: http://journals.cambridge.org/action/displayFulltext?type=6&fid=2818204&jid=CTY&volumeId=18&issueId=S2&aid=2818200&bodyId=&membershipNumber=&societyETOCSession=&fulltextType=RA&fileId=S1047951108002916

6. Temporary Certification Program for Access to the Death Master File [Internet]. [cited 2014 Jul 10]. Available from: https://www.federalregister.gov/articles/2014/03/26/2014-06701/temporary-certification-program-for-access-to-the-death-master-file

7. Data Access - National Death Index. [cited 2014 Jul 2]; Available from: http://www.cdc.gov/nchs/ndi.htm

8. Hauser T. Accuracy of on-line databases in determining vital status. J Clin Epidemiol [Internet]. 2001 Dec [cited 2014 Jul 1];54(12):1267–70. Available from: http://www.sciencedirect.com/science/article/pii/S0895435601004218

9. 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 [Internet]. 2013 Jan [cited 2014 Jul 28];3(5):518–22. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3816971&tool=pmcentrez&rendertype=abstract

10. Cowper DC, Kubal JD, Maynard C, Hynes DM. A Primer and Comparative Review of Major U.S. Mortality Databases. Ann Epidemiol [Internet]. 2002 Oct [cited 2014 Jul 1];12(7):462–8. Available from: http://www.sciencedirect.com/science/article/pii/S104727970100285X

11. Hill ME, Rosenwaike I. Social Security Administration's Death Master File: The Completeness of Death Reporting at Older Ages, The. Soc Secur Bull [Internet]. 2001 [cited 2014 Jul 7];64. Available from: http://heinonline.org/HOL/Page?handle=hein.journals/ssbul64&id=49&div=&collection=journals

12. Schisterman EF, Whitcomb BW. Use of the Social Security Administration Death Master File for ascertainment of mortality status. Popul Health Metr [Internet]. 2004 Mar 5 [cited 2014 Jul 2];2(1):2. Available from: http://www.pophealthmetrics.com/content/2/1/2

13. Sesso HD, Paffenbarger RS, Lee I-M. Comparison of National Death Index and World Wide Web Death Searches. Am J Epidemiol [Internet]. 2000 Jul 15 [cited 2014 May 31];152(2):107–11. Available from: http://aje.oxfordjournals.org/content/152/2/107.full

14. Internet Scientific Publications [Internet]. [cited 2014 Jul 2]. Available from: http://ispub.com/IJE/11/2/1604

15. Myers RB, Herskovic JR. Probabilistic techniques for obtaining accurate patient counts in Clinical Data Warehouses. J Biomed Inform [Internet]. 2011 Dec [cited 2014 Jul 22];44 Suppl 1:S69–77. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3251720&tool=pmcentrez&rendertype=abstract